Investigating Domain Adaptation in Sentiment Analysis: A Comparative Study on IMDB, Amazon, and Twitter(X)

Anonymous ACL submission

Abstract

Cross-domain Sentiment Analysis (CD-SA) presents a formidable challenge due to the domain shift between training and test data. Traditional sentiment classification models often exhibit performance degradation when applied to data from unfamiliar domains. In this study, we establish a robust baseline using Sentence-BERT (SBERT) and investigate multiple optimization strategies to improve cross-domain adaptability.

We evaluated various feature representations, including TF-IDF, Word2Vec, and FastText, in conjunction with various classifiers such as Logistic Regression, LSTM, and XGBoost. Furthermore, we explore the impact of text preprocessing techniques, including spelling correction, slang normalization, and emoji translation, on classification performance. Our findings demonstrate that a combination of TF-IDF and FastText enhances sentiment classification accuracy, while pre-trained BERT achieves the highest performance on Twitter data but incurs computational costs. Additionally, we observe that excessive text preprocessing does not necessarily lead to improved model performance and may introduce extraneous information.

Our findings underscore the significance of striking a balance between feature engineering and model complexity in CD-SA. We provide valuable insights into effective strategies for cross-domain sentiment classification and elucidate the trade-offs between computational efficiency and classification accuracy.

1 Introduction

Sentiment Analysis (SA) has gained significant attention in Natural Language Processing (NLP) due to its widespread applications in customer feedback analysis, product reviews, and social media monitoring. While traditional supervised sentiment classification models perform well within the same domain, they often suffer from substantial performance degradation when applied to unseen domains (Blitzer et al., 2007). This challenge, known as Cross-domain Sentiment Analysis (CD-SA), arises due to domain shifts in vocabulary, writing style, and sentiment expression across different datasets.

Existing approaches to CD-SA typically rely on feature engineering, domain adaptation techniques, or pre-trained language models. Traditional methods such as TF-IDF combined with Logistic Regression (LR) provide robust performance within a single domain but lack generalizability to new domains (Wang and Manning, 2012). Deep learning models, such as Long Short-Term Memory (LSTM) networks and Transformer-based architectures, have shown improvements in feature representation learning. However, their performance in CD-SA is still largely affected by domain-specific variations (Howard and Ruder, 2018). Moreover, excessive text preprocessing, such as spelling correction, slang normalization, and emoji translation, may not always lead to improved model performance and can sometimes introduce noise.

In this study, we aim to investigate the effectiveness of various sentiment classification methods in CD-SA. We establish a strong baseline using Sentence-BERT (SBERT) and evaluate different feature representations, including TF-IDF, Word2Vec, and FastText, in combination with multiple classifiers such as Logistic Regression, LSTM, and XGBoost. Additionally, we examine the impact of various text preprocessing techniques on cross-domain adaptability.

Our key contributions are summarized as follows:

- We introduce SBERT as a cross-domain baseline and compare its performance with traditional and deep learning-based sentiment classification methods.
- We analyze the effectiveness of text prepro-

cessing techniques, such as spelling correction, network slang normalization, and emoji translation, in enhancing CD-SA.

- We demonstrate that a combination of TF-IDF and FastText improves sentiment classification accuracy, while pre-trained BERT achieves the highest performance (77%) on Twitter data but at a high computational cost.

- We provide insights into the trade-offs between feature engineering and model complexity in CD-SA and highlight the limitations of excessive text preprocessing.

Our experimental results on three datasets (IMDB, Amazon, and Twitter) provide a comprehensive evaluation of cross-domain sentiment classification methods. The findings from this study contribute to a deeper understanding of how different representation learning strategies impact CD-SA and provide guidance for future research on domain adaptation in sentiment analysis.

2 Related Work

2.1 Sentiment Analysis Methods

Sentiment Analysis (SA) is a fundamental task in Natural Language Processing (NLP) that aims to classify the sentiment polarity of text (Pang and Lee, 2008). Traditional approaches rely on statistical and rule-based methods, such as Term Frequency-Inverse Document Frequency (TF-IDF) combined with Logistic Regression (LR) (Wang and Manning, 2012). While these methods perform well within the same domain, their generalizability across domains is often limited due to vocabulary shifts and domain-specific expressions.

Deep learning-based approaches have shown significant improvements in sentiment classification by leveraging word embeddings and neural architectures. Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017) provide dense vector representations that capture semantic relationships between words. Long Short-Term Memory (LSTM) networks (Howard and Ruder, 2018) and Transformer-based models, such as BERT (Devlin et al., 2019), have further enhanced feature extraction by capturing contextual dependencies in text. However, these models still struggle with domain adaptation, as they are primarily trained on in-domain datasets.

2.2 Cross-Domain Sentiment Analysis

Cross-Domain Sentiment Analysis (CD-SA) aims to address the generalization challenge in sentiment classification by transferring knowledge from a source domain to a target domain (Blitzer et al., 2007). One common approach is domain adaptation, where models are trained to align feature distributions between different domains. Representation learning techniques, such as adversarial training and contrastive learning, have been explored to mitigate domain shifts (Howard and Ruder, 2018).

Pre-trained language models, such as BERT and SBERT (Reimers and Gurevych, 2019), have shown promising results in CD-SA by leveraging large-scale textual knowledge. However, fine-tuning such models requires significant computational resources, making them less practical for real-world applications. Additionally, excessive fine-tuning may lead to overfitting on the source domain, reducing the model's adaptability to new domains.

2.3 The Impact of Text Preprocessing in Sentiment Analysis

Text preprocessing plays a crucial role in sentiment analysis by removing noise and standardizing text inputs. Common preprocessing steps include spelling correction, slang normalization, and emoji translation (Pang and Lee, 2008). While these techniques are beneficial for structured datasets, they may not always improve performance in CD-SA. Excessive preprocessing can introduce noise, especially in social media data where informal expressions contribute to sentiment cues.

Our study systematically evaluates the impact of text preprocessing on CD-SA by comparing different levels of preprocessing and their effects on model performance. We demonstrate that balancing preprocessing and model complexity is essential for achieving robust cross-domain sentiment classification.

2.4 Our Contributions

While existing studies have explored various methods for CD-SA, our work differs in several key aspects:

- We evaluate multiple feature representations, including TF-IDF, Word2Vec, FastText, and SBERT, in the context of CD-SA.

- We analyze the impact of different text preprocessing techniques on model performance

across multiple datasets.

- We investigate the effectiveness of hyperparameter optimization using Optuna for models such as XGBoost and LightGBM.

- We provide insights into the trade-offs between feature engineering, model complexity, and computational efficiency in CD-SA.

3 Datasets

3.1 Dataset Description

To evaluate the effectiveness of different sentiment classification methods in cross-domain sentiment analysis (CD-SA), we use three datasets from distinct domains: **IMDB Reviews**, **Amazon Reviews**, and **Twitter (X)**.

- **IMDB Reviews**: This data set consists of 50,000 movie reviews labeled positive or negative (Pathi, 2021). Reviews are relatively well-structured and contain longer text segments compared to social media posts. The data set is publicly available on Kaggle [1].

- **Amazon Reviews**: This data set includes 400,000 product reviews that span multiple categories such as electronics, clothing and household items (Chavan, 2021). The sentiment labels are derived from user ratings, where reviews with 1-2 stars are labeled as negative, 4-5 stars as positive, and 3-star reviews are excluded. The data set is available on Kaggle [2].

- **Twitter (X)**: This dataset, also known as Sentiment140, originally contained 1,600,000 tweets collected from various Twitter users (Sentiment140, 2009). To balance computation and maintain diversity, we randomly sampled 400,000 tweets. Unlike IMDB and Amazon reviews, Twitter tweets are short, informal, and contain slang, emojis, and abbreviations, making them challenging for standard NLP models. The data set can be accessed on Kaggle [3].

[1] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2] https://www.kaggle.com/datasets/jagdishchavan/amazon-reviews

[3] https://www.kaggle.com/datasets/kazanova/sentiment140

3.2 Data Preprocessing

To ensure consistency between datasets and improve model performance, we applied a series of pre-processing steps:

- **Tokenization**: We experimented with both **NLTK** and **SpaCy** for tokenization. While SpaCy provides efficient processing, NLTK allows for more fine-grained customization.

- **Stopword Removal**: Common stopwords (e.g., "the", "is", "and") were removed to reduce noise.

- **Lowercasing**: All text was converted to lowercase to maintain uniformity.

- **Punctuation and Special Character Removal**: HTML tags, URLs, and special symbols were eliminated.

- **Text Normalization**: For the Twitter dataset, we performed additional preprocessing:
  - **Spelling Correction**: Fixed common misspellings using an NLP-based spell checker.
  - **Slang Normalization**: Converted abbreviations (e.g., "luv" → "love", "idk" → "I don't know").
  - **Emoji Translation**: Mapped emojis to corresponding textual descriptions.

3.3 Data Analysis

To better understand the characteristics of each dataset, we performed exploratory data analysis (EDA), including text length distribution and sentiment class balance.

Dataset	Samples	Avg. Text Length	Sentiment Balance (%)
IMDB	49,582	230.1	50 (pos) / 50 (neg)
Amazon	400,000	120.5	52 (pos) / 48 (neg)
Twitter (X)	400,000	28.3	50 (pos) / 50 (neg)

Table 1: Dataset statistics: number of samples, average text length, and sentiment class balance.

The table shows that IMDB reviews have significantly longer texts than Amazon reviews and tweets. The Twitter dataset contains extremely short messages, which makes feature extraction challenging.

These variations in dataset characteristics highlight the necessity of evaluating different feature representations and classification methods in CD-SA.

4 Methodology

4.1 Problem Definition

Cross-Domain Sentiment Analysis (CD-SA) aims to classify the sentiment polarity of textual data while ensuring that the model generalizes well across different domains. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a textual review and $y_i \in \{0, 1\}$ represents the sentiment label (negative or positive), our goal is to learn a sentiment classifier $f : x \rightarrow y$ that maintains high accuracy even when tested on data from a domain different from the training domain.

4.2 Evaluation Metrics

To evaluate model performance, we use the following metrics:

- **Accuracy**: Measures the overall percentage of correctly classified samples.
- **F1-score**: Computes the harmonic mean of precision and recall, providing a balanced evaluation for imbalanced datasets.

4.3 Baseline Model

To establish a reference for performance comparison, we evaluate a widely used cross-domain sentiment classification approach.

SBERT (all-MiniLM-L6-v2) + Logistic Regression We use Sentence-BERT (SBERT) with the all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019) to generate sentence embeddings for Amazon reviews. These embeddings are then used to train a Logistic Regression classifier, which is evaluated on the IMDB dataset. SBERT embeddings provide strong semantic representations, making them a reliable baseline for cross-domain generalization.

4.4 Proposed Methods

To improve CD-SA performance, we introduce several optimization strategies.

4.4.1 Optimizations for Amazon → IMDB

Since IMDB reviews are well-structured but differ in style from Amazon product reviews, we explored the following improvements:

TF-IDF + Logistic Regression To enhance IMDB classification performance, we trained a TF-IDF + Logistic Regression model on Amazon data. This method significantly improved accuracy on IMDB compared to SBERT + Logistic Regression, demonstrating that simple statistical features can be effective in cross-domain adaptation.

Switching to SpaCy for Tokenization Instead of using NLTK for tokenization, we experimented with SpaCy, which provides more efficient tokenization and built-in support for named entity recognition (NER) and part-of-speech (POS) tagging. This change improved preprocessing speed while maintaining comparable accuracy.

Word2Vec + LSTM We trained a Word2Vec model on the Amazon dataset and applied an LSTM network to classify IMDB reviews. The results were comparable to TF-IDF + Logistic Regression, but LSTM provided better flexibility in handling longer text dependencies.

4.4.2 Optimizations for Amazon → Twitter

Since Twitter data contains slang, emojis, and abbreviations, we implemented a series of optimizations:

TF-IDF + Logistic Regression on X We tested the TF-IDF + Logistic Regression model, trained on Amazon, on Twitter data. Unlike IMDB, this model performed poorly on Twitter, likely due to the informal and highly variable language style of tweets.

Customized Tokenization for Twitter Data To better handle informal text, we applied:

- **Normalizing informal expressions**: (e.g., "u" → "you")
- **Expanding abbreviations**: (e.g., "omg" → "oh my god")
- **Mapping emojis to sentiment-related text equivalents**

Despite these optimizations, performance gains were minimal.

Hyperparameter Optimization on X Data We applied hyperparameter tuning using Optuna (Akiba et al., 2019) on:

- **XGBoost**: Tuned learning rate, max depth, and feature selection.
- **LightGBM**: Adjusted tree-based feature learning parameters.

However, none of these models significantly outperformed the baseline.

Fine-Tuning Twitter-RoBERTa To further enhance performance on Twitter data, we tested the 'twitter-roberta-base-sentiment' model (Barbieri et al., 2020). Unlike traditional BERT-based approaches, this model is pre-trained on social media data, making it highly effective for sentiment classification in the Twitter domain. It significantly outperformed other models on X, demonstrating the advantage of domain-specific pretraining.

5 Experiments

5.1 Experimental Setup

To evaluate the effectiveness of different sentiment classification methods in Cross-Domain Sentiment Analysis (CD-SA), we train models on the **Amazon Reviews** dataset and test them on two unseen datasets: **IMDB** and **Twitter (X)**. This setup allows us to assess the model's ability to generalize across different domains.

Hyperparameter Settings For traditional models (TF-IDF + Logistic Regression, XGBoost, LightGBM), we apply hyperparameter tuning using Optuna (Akiba et al., 2019) to select the optimal learning rate, regularization strength, and tree-based model parameters. The LSTM model is optimized with respect to hidden size, dropout rate, and training epochs. Fine-tuning of Twitter-RoBERTa follows best practices for transformer-based models, including optimal batch size and learning rate adjustments.

5.2 Baseline vs. Proposed Models

We evaluate the performance of the following models:

- **Baseline Model:** SBERT (all-MiniLM-L6-v2) + Logistic Regression.
- **Optimized Models:** TF-IDF + Logistic Regression, Word2Vec + LSTM, XGBoost with Optuna hyperparameter optimization, LightGBM, Twitter-RoBERTa ('twitter-roberta-base-sentiment').

5.3 Evaluation Metrics

To comprehensively assess model performance, we report the following metrics for all experiments:

- **Accuracy**: Used as a general performance measure for all models, particularly in structured text settings such as IMDB reviews.
- **F1-score**: Given the class imbalance in the Twitter dataset, we prioritize F1-score over accuracy for evaluating models, as it provides a better balance between precision and recall.
- **Per-Class Analysis**: For the Twitter dataset, we report separate F1-scores for positive and negative sentiment, highlighting potential biases in different models.

Unlike IMDB, where accuracy is a reliable indicator of performance, we find that accuracy alone is insufficient for Twitter sentiment analysis due to class imbalance and informal language variations. Therefore, we rely primarily on F1-score and per-class sentiment analysis to evaluate models on the Twitter dataset.

5.4 Experimental Results

Table 2 presents the performance of different models, where all models are trained on **Amazon Reviews** and tested on IMDB and Twitter (X).

Model	Amazon → IMDB	Amazon → Twitter
SBERT (all-MiniLM-L6-v2) + LR	0.77	0.68
TF-IDF + LR	**0.87**	0.61
Word2Vec + LSTM	0.85	-
XGBoost (Optuna)	-	0.57
LightGBM	-	0.57
Twitter-RoBERTa	-	**0.77**

Table 2: Experimental results: accuracy of models trained on Amazon Reviews and tested on IMDB and Twitter.

5.5 Analysis of X Dataset

To further analyze the challenges of cross-domain sentiment classification on Twitter (X), we compare different models and preprocessing techniques based on their F1-score for positive and negative sentiments.

Model	Negative F1-score	Positive F1-score
TF-IDF + LR	0.63	0.58
XGBoost (Optuna)	0.43	0.65
LightGBM	0.38	0.67
Twitter-RoBERTa	**0.76**	**0.78**

Table 3: Performance of different models on X dataset with F1-score for positive and negative sentiment.

Results in Table 3 show that domain-specific models such as Twitter-RoBERTa significantly improve sentiment classification on Twitter, achieving the highest F1-scores for both negative (0.76) and positive (0.78) sentiments. This demonstrates that models trained on social media data are better suited for handling informal language.

Traditional models like TF-IDF + Logistic Regression showed strong performance in detecting negative sentiment (0.63 F1-score) but struggled with positive sentiment (0.58 F1-score). This suggests that rule-based term frequency methods may be more effective at capturing explicit negative words but less robust for informal positive sentiment.

On the other hand, XGBoost and LightGBM, despite hyperparameter tuning, showed inconsistent performance. XGBoost achieved relatively high positive sentiment detection (0.65 F1-score) but had poor recall for negative sentiment (0.43 F1-score), indicating strong bias towards positive classification. LightGBM performed even worse on negative samples (0.38 F1-score) but slightly better on positive sentiment (0.67 F1-score).

Overall, these results highlight that while traditional models and tree-based classifiers struggle with the complexity of Twitter language, domain-adapted transformers such as Twitter-RoBERTa significantly outperform other methods.

6 Conclusion

In this study, we explored various methods for Cross-Domain Sentiment Analysis (CD-SA), evaluating different feature representations and classification models trained on the Amazon dataset and tested on IMDB and Twitter (X).

6.1 Summary of Contributions

Our research provides key insights into the effectiveness of different approaches for CD-SA:

- We established SBERT (all-MiniLM-L6-v2) + Logistic Regression as a strong baseline for cross-domain generalization.
- We demonstrated that TF-IDF + Logistic Regression significantly improves performance on IMDB but struggles to generalize to Twitter due to informal language variations.
- We evaluated tree-based models (XGBoost, LightGBM) with hyperparameter tuning, showing that they failed to significantly outperform simple statistical models on the Twitter dataset.
- We confirmed that Twitter-RoBERTa, a domain-specific pre-trained model, outperforms all other approaches on Twitter, highlighting the importance of adapting models to domain-specific linguistic patterns.

6.2 Challenges and Limitations

During our research, we identified several challenges and limitations in CD-SA:

Text Processing Complexity We found that excessively detailed text preprocessing does not always improve model performance and may even have negative effects. For example, while slang normalization and emoji translation provided slight improvements, overly complex text modifications sometimes introduced noise rather than useful information.

Cross-Domain Adaptation in Twitter Data Despite applying various fine-grained processing techniques on Twitter data, we observed that these efforts did not significantly improve classification accuracy. Surprisingly, TF-IDF + Logistic Regression performed comparably to more advanced models. This suggests that the linguistic style of Twitter is too diverse, making traditional text-processing pipelines less effective.

Computational Constraints We initially aimed to integrate advanced spelling correction and typo normalization, but processing time was too high, especially for longer texts. As a compromise, we limited typo correction to short phrases to maintain efficiency.

Model Training Limitations Due to hardware constraints, we were unable to fine-tune BERT-based models on our dataset. If computational resources were available, we believe fine-tuning BERT or other large-scale transformers could further improve CD-SA performance.

6.3 Key Findings

Experimental results indicate that traditional models (e.g., TF-IDF + Logistic Regression) can be effective for structured text such as IMDB reviews but are limited when dealing with informal text like tweets. Despite hyperparameter tuning, XGBoost and LightGBM exhibited significant class imbalance issues when applied to Twitter. In contrast, Twitter-RoBERTa achieved the highest F1-scores for both positive and negative sentiment classification, demonstrating the advantages of pretraining on domain-specific data.

Additionally, our findings suggest that preserving essential textual information is crucial in cross-domain sentiment analysis. Instead of aggressively filtering or modifying text, models must learn to

extract meaningful sentiment cues while ignoring noise.This balance is particularly important when handling informal language sources like Twitter.

6.4 Future Work

For future research, we propose the following directions:

- Improving Cross-Domain Adaptation: Exploring advanced transfer learning techniques such as adversarial domain adaptation or contrastive learning to enhance CD-SA performance.
- Leveraging More Pretrained Models: Investigating the impact of large-scale transformers like GPT-4 and T5 on CD-SA.
- Efficient Training Strategies: If computational resources become available, fine-tuning BERT-based models for cross-domain sentiment analysis should be explored.
- Refining Text Processing Strategies: Developing methods to preserve important sentiment-related features while minimizing unnecessary modifications.

Overall, our study highlights the challenges and opportunities in CD-SA, demonstrating that while traditional models provide a strong baseline, domain-specific transformer models such as Twitter-RoBERTa are crucial for achieving state-of-the-art performance on informal text sources.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1644–1650.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *ACL*, pages 440–447.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jagdish Chavan. 2021. Amazon reviews dataset.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, pages 4171–4186.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *ACL*, pages 328–339.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Lakshmi Narayan Pathi. 2021. Imdb dataset of 50k movie reviews.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.

Sentiment140. 2009. Sentiment140 - twitter sentiment analysis dataset.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. *ACL*, pages 90–94.